IMPROVING AUDIO WATERMARKING SCHEME USING PSYCHOACOUSTIC WATERMARK FILTERING

Nedeljko Cvejic, Tapio Seppanen

MediaTeam
Information processing laboratory
FIN-90014 University of Oulu, Finland

ABSTRACT

A novel algorithm for embedding a spread-spectrum-based watermark into uncompressed, raw audio sequences is presented. The scheme efficiently takes advantage of masking phenomena in HAS in order to embed watermark data below the masking threshold of audio signal. Detection of the watermark is done by blind detection, without using the original audio. None of the transformations to and from frequency domain are performed either in embedding or extraction part of the proposed scheme, resulting in the calculation simplicity of the embedding and detection process. In experimental tests, the scheme proved to be robust against common attacks against audio watermarking algorithms. Subjective quality evaluation of the algorithm showed that embedded watermark introduces low, inaudible distortion of host audio signal.

1. INTRODUCTION

Today’s multimedia systems’ properties raise the question of copyright protection, monitoring of the broadcast signals, and development of algorithms and real-time applications that would disable illegal copying and redistribution of previously pirated material. One of the possible solutions in that area is data watermark, which is added to multimedia content by embedding an imperceptible and statistically undetectable signature. Thereby, multimedia data creators and distributors are able to prove ownership of intellectual property rights without forbidding other individuals to copy the multimedia content itself. Watermarking techniques were primarily developed for digital images and video sequences; interest and research in audio watermarking started slightly later.

However, in the last five years, several algorithms for embedding and extraction of watermarks in audio sequences have been presented [1-6]. All of the cited algorithms take advantage of perceptual properties of the human auditory system (HAS), foremost its “vulnerability” to masking in the frequency and temporal domain, in order to add additional bits into a host signal in a perceptually transparent manner. Embedding of additional bits in audio signal is a considerably more tedious task than implementation of the same process on images and video, due to the dynamical superiority of HAS in comparison with the human visual system. As stated in [6], HAS receives information over a range of power of one billion to one and a range of frequencies greater than thousand to one. On the other hand, besides masking in temporal and frequency domain, HAS is insensitive to a constant relative phase change in a static waveform and some specific spectral distortions interpret as natural.

Recently, we developed a robust audio watermarking algorithm for copyright protection of digital audio [8]. The procedure uses a time domain embedding algorithm and properties of spread spectrum communications as well as temporal and frequency-domain masking in HAS. None of the generally employed transformations to and from frequency domain are performed, either on embedding or extraction part of the proposed scheme. In the present paper, we improve the performance of our method by utilizing more of the HAS properties in the watermark embedding algorithm. The basic idea is that the spectrum of the m-sequence is shaped in accordance to the HAS in order to make the watermark even more imperceptible. We add an integration function with a synchronization scheme in the receiver for better attack resistance and decrease of the computational complexity of the extraction algorithm. For handling time scaling attacks, a multiple chip embedding is used. With these enhancements, we achieve a considerably lower demand for computational power, blind watermark detection and better time scaling resistance than with our earlier algorithm. The experiments indicated excellent robustness to many watermark attacking schemes and good performance in the presence of mp3 and AAC compression as well.

2. WATERMARK EMBEDDING

Figure 1 gives a general overview of the watermark-embedding algorithm. The embedding scheme proposed in this paper modifies the original audio signal, which is represented as a 16-bit sample sequence sampled at 44100 Hz, mono. The m-sequence is obtained from a shift register with feedback and represented in the bipolar form {-1,1}. Prior to further processing, the m-sequence is filtered in order to adjust it to masking thresholds of the human auditory system (HAS) in the frequency domain. The main goal is to adapt the watermark to such form that the energy of the watermark is maximized under the restriction of keeping auditory distortions to a minimum, although the SNR value is significantly decreased [see Table 1]. The frequency characteristic of the filter is the approximation of the threshold in quiet curve of the HAS, plotted in Figure 2. Despite the simplicity of the shaping process of the m-sequence in frequency domain, the result is an inaudible watermark as the largest amounts of the shaped watermark's power are concentrated in the frequency sub-bands with lower HAS sensitivity. In addition, these frequency sub-bands (frequencies below 500 Hz and above 11000Hz) are an essential part of the watermarked audio and cannot be removed from its spectrum without making serious damage to the perceptual quality. A significant number of computational operations needed for frequency analysis of audio, which have to be run in order to derive global masking thresholds in a predefined time window, are skipped, making this scheme appreciably faster. Although standard frequency analyses have more accurate data about the audio spectrum, simulation tests done with selected audio clips showed a high level of similarity with the frequency masking thresholds derived from the masking model defined in ISO-MPEG Audio Psychoacoustic Model [7]. Parameter α can always be set to the value which places the masking curve of the algorithm near or under the most stringent local threshold value defined by standard masking model.

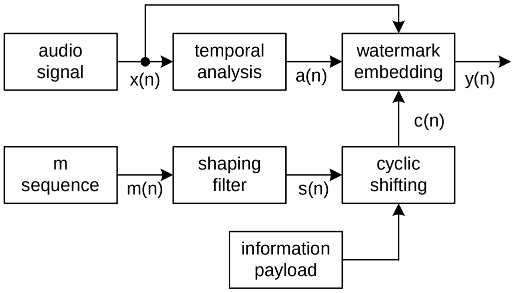

Figure 1: Watermark embedding scheme

A cyclic shifted version c(n) is used to achieve a multi-bit payload for one particular watermark sequence s(n). Every possible shift may be associated with different information content. Therefore, information payload is directly proportional to the length of the m-sequence. The cyclical shifting of the shape-filtered m-sequence changes only the phase but not the amplitude of the spectrum; therefore, the desired spectral shaping is retained. There is always a possibility to make the trade-off between the embedded data size and robustness of the algorithm [4]; as the m-sequence length is decreasing, the algorithm is able to add more bits into the host audio but the detection of the hidden bits and resistance to different attacks is decreased. We found that an m-sequence length of 1023 samples is a good compromise between the amount of hidden content and the algorithm's attack resistance.

Host audio sequence is also analysed in the time domain, where a minimum or a maximum is determined in the block of audio signal that has the length of 7.6 ms. The goal of the temporal analysis is to place the watermark inside the raw audio without making any perceptual distortion to the host signal by using temporal masking characteristics of the HAS. The masker value determined from the temporal analysis is used as a reference point to determine the level of power of the watermark sequence in the analysed block. With reference to temporal masking curves and the length of analysed audio frames, it was concluded that the added information should be at least 24 dB below the power level of the audio maximum in the frame. The algorithm equally uses both pre- and post-masking properties, therefore making the most significant error if the maximum of the host audio is situated at the end of the analysed block. However, the impact of sub-maximums and the maskers from the contiguous blocks is not negligible and it helps the current masker to fulfil its role. As the result of this analysis, the watermark samples are weighted by the coefficient a(n) in order to be adjusted to psycho-acoustic perceptual thresholds.

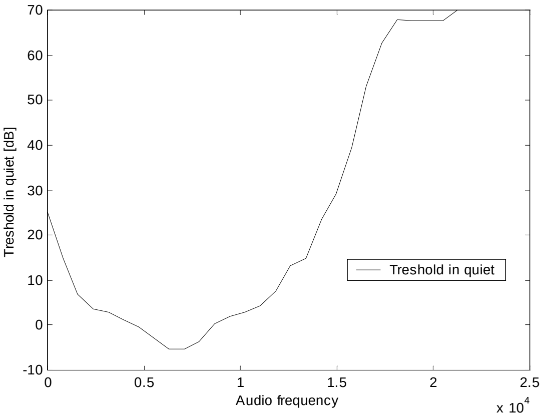

Figure 2: Threshold in quiet curve of HAS

Therefore, the watermark signal is embedded into host audio using three time-aligned processes. In the first stage, the m-sequence has been filtered with the shaping filter, where a coloured-noise sequence s(n) is the output. Samples of the s(n) sequence are then cyclically shifted, where the shift value is dependant of the input information payload. At the output of the watermark embedding scheme, shifted version of s(n), sequence c(n) is being weighted and embedded to the original audio signal:

$$y(n) = x(n) + \alpha \cdot a(n) \cdot c(n)$$

where x(n) denotes input audio signal, a(n) are coefficients from the temporal analysis block and α is a parameter that represents the trade-off between perceptual transparency and detection reliability. However, addition of the c(n) sequence in the embedding process is done repeatedly in order to make the system resistant to time scaling attacks that tend to de-synchronize the extraction process (see watermark extraction). As α increases, robustness of the embedded watermark is better, but it is limited by allowable perceptual distortion of the watermarked audio. Subjective listening tests were performed on different audio clips in order to experimentally determine the maximum value for α. High perceptual transparency was achieved for $\alpha \in (0.1, 0.4)$, depending on the type of music.

3. WATERMARK EXTRACTION

The diagram of the audio watermark detection scheme is shown in Figure 3. The proposed detection procedure does not require access to the original signal to detect the embedded watermark. The cornerstone of detection process is the mean removed cross-correlation between the watermarked audio signal and the equalized m-sequence. If we define y*(n) as the watermarked audio samples and m(n) as equalized m-sequence samples, the raw cross-correlation values $c_{my}(m)$ are defined as:

$$c_{my}(m) = \begin{cases} \sum_{n=0}^{N-|m|-1} \left(m(n) - \frac{1}{N} \sum_{i=0}^{N-1} m_i \right) \left(y^*_{n+m} - \frac{1}{N} \sum_{i=0}^{N-1} y^*_i \right), m \geq 0 \\ c^*_{ym}(-m), m < 0 \end{cases}$$

However, before the watermarked signal is segmented into blocks in order to measure the cross-correlation with the m-sequence, the detection algorithm filters it with the equalization filter [5]:

$$y(n) = y^*(n) * d(n);$$
$$d(n) = [-1\ 2\ -1]$$

Equalization is also performed on the m-sequence in order to match the incoming watermark samples as precisely as possible, regarding known modification (equalization) of watermarked audio. Generally, in a correlation detector scheme it is often assumed that the communication channel is white Gaussian. Nevertheless, statistics for real audio signals show that audio samples are highly correlated. Applying a whitening procedure should considerably reduce any correlation in the audio and thus achieve optimum detection. The equalization filter suppresses the low-frequency components with high energy and emphasizes the high-frequency part of the audio spectrum in order to obtain a more flat, “noise-like” spectrum. Frequency domain shaping process of m-sequences performed in the embedding part of the scheme is repeated in the extraction as well, in order to optimise matched filtering performance. Hence, input sequences of the auto-correlation process are m-sequences with the same temporal and frequency domain characteristics, if there is no signal processing of the watermark and detection is synchronized. Before the start of the integration process, which determines the peak and the output value, the block power normalization part of the scheme makes uniform energies of the output blocks from correlation calculations. Thereby every output block (length 2045 samples) from the correlator has an approximately equal impact on the integration process that lasts for 84 consecutive frames. Otherwise, a “poor“ correlation result in a block with high amplitudes of the host audio diminishes the result of the many positive correlation detections, which could often happen in very dynamic signals as audio signal. The integration block sums the normalized output block from correlation detection and determines the peak and its position. The peak value is related directly to the detection reliability, whereas its position corresponds to the cyclic shift – information payload. The detection reliability depends strongly on the number of accumulated frames. In general, the trade-off is made between the time of integration and the amount of hidden data.

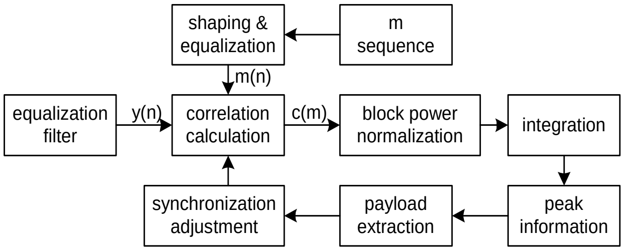

Figure 3: Watermark extraction scheme

The correlation method and the watermark extraction algorithm in general are reliable only if correlation frames are aligned with those used in watermark embedding. Therefore, one of the malicious attacks can be de-synchronization of the cross-correlation procedure by time-scale modifications. In that case, the watermark detection scheme is not properly determining the shift value in the embedded watermark, resulting in high increase of the BER. One of the methods against time scaling, which has been chosen for this watermark extraction scheme, is to use redundancy in the watermark chip pattern, similar to one described in [3]. The basic idea is to spread each chip of the shaped m-sequence onto R consecutive samples of watermarked audio. It has been proved that, using such an embedding and detection scheme, the correlation is correctly calculated even if a linear shift of floor(R/2) samples across the temporal or frequency domain is induced. However, there is a trade-off between robustness of the algorithm and computational complexity, which is significantly increased by performing multiple correlation tests.

4. EXPERIMENTAL RESULTS

Subjective quality evaluation of the watermarking method has been done by listening tests involving ten persons. A total number of eight audio pieces were used as tests signals, of 10 s duration each. The audio excerpts were selected so that they represent a broad range of music genres, i.e. audio clips with different dynamic and spectral characteristics. In the first part of the test, participants listened to the original and the watermarked audio sequences and were asked to report dissimilarities between the two signals, using a 5-point impairment scale: (5: imperceptible, 4: perceptible but not annoying, 3: slightly annoying, 2:annoying 1: very annoying). Table 1 presents results of the first test, with the lowest and the highest value from the impairment scale and average MOS for given audio excerpt. In the second part, test participants were repeatedly presented with unwatermarked and watermarked audio clips and were asked to determine which one is the watermarked one. Experimental results are presented also in Table 1, values near to 50% show that the two audio clips (original audio sequence and watermarked audio signal) cannot be discriminated.

Table 1: SNR, mean opinion scores and discrimination of original and watermarked audio excerpts (α=0.3)

file name	SNR	discrimination	MOS range	average MOS
lovett	18	53%	4-5	4.6
ritenour	18.2	49%	4-5	4.8
yoyoma	19.2	50%	5	5
titanic	18.1	52%	4-5	4.8
yanni	17.7	47%	4-5	4.8
joecocker	16.4	49%	4-5	4.5
abba	16.4	54%	3-5	4.2
eurythmics	17.4	45%	3-5	4.1
total average MOS				4.56

In order to evaluate the algorithm's resistance to common attacks, the sequences were tested against mp3 and AAC coding, and a set of other modifying functions. The audio segments used in experiments were mono signals, 20 seconds long, sampled at 44.1 KHz with 16 bits resolution. Processing was performed in MatLab and CoolEdit 2000. Additional 100 bits were embedded in each of the eight audio sequences; the time of integration in watermark extraction scheme was set to two seconds, resulting in a watermark bit rate of 5b/s. The audio clips were compressed to MPEG layer-3 files, at a rate of 32 kb/s using Syntrillium's commercial mp3 coder based on software implementation licensed from the Fraunhofer IIS. The extraction results after the employed compression are presented in Table 2. A similar test was executed for the AAC compression. The audio sequences were encoded with Freeware Advanced Audio Coder/Decoder at the rate of 48kb/s. The detection performance of the algorithm was also tested against common signal processing attacks [5]:

1. All-pass filtering using system function: $H(z)=(0.81z^2 - 1.64z + 1) / (z^2 - 1.64z + 0.81)$
2. Echo-addition (delay 100ms, decay 50%)
3. Band-pass filtering using a second order Butterworth filter with cut-off frequencies 100 Hz and 3000 Hz
4. Amplitude compression (8.91:1 for A>-29dB, 1.73:1 for –46dB<A<-29dB and 1:1.61 for A<-46dB)
5. Equalization (6-band equalizer, signal suppressed or amplified by 6 dB in each band)
6. Noise addition (with uniform white noise. Maximum magnitude of 200 quantization steps)
7. Time-scale modification of –4% or +4%, where the pitch remains unaffected.

Detection results for the various attacks described above are shown in the Table 2. Columns beneath the attack description list the number of incorrectly detected bits (of 100 bits embedded in total) after the attack was performed. The algorithm obtained a perfect detection result in the cases of equalization, all-pass filtering, amplitude compression, echo addition, and noise addition. In the attacks done by mp3 and AAC compression and time-scaling, the bit error rate is higher than in the case of other attacks, but detection performance is still within an acceptable range. In general, experimental tests showed that the watermark was embedded in accordance with HAS properties, using temporal and frequency domain masking techniques yielding good average MOS and low discrimination between the original and watermarked audio sequence. The algorithm has high detection results, after different signal processing and compression attacks, with the worst results after mp3 and AAC coding and time-scaling.

Table 2: Extraction results after attacks (α=0.3)

Watermark attack type/ filename	mp3	AAC	equalization	all-pass	echo	low- pass	amp.compr.	time -4%	time +4%	noise
lovett	4	0	0	0	0	0	0	3	5	0
ritenour	0	0	0	0	0	0	0	0	0	0
yoyoma	0	0	0	0	0	0	0	0	0	0
titanic	0	0	0	0	0	2	0	5	3	0
yanni	0	0	0	0	0	0	0	0	0	0
joecocker	3	1	0	0	0	0	0	0	0	0
abba	3	5	0	0	0	0	0	1	3	0
eurythmics	3	5	0	0	0	0	0	2	1	0
total	13	11	0	0	0	2	0	11	12	0
mean BER (%)	1.62	1.38	0	0	0	0.25	0	1.38	1.5	0

The reason for poorer extraction capabilities after mp3 and AAC coding is that these compression techniques crop high frequency spectrum of the watermarked audio, where most of the watermark energy is situated. Time scaling (stretching, DA/AD conversion, etc.) is always one of the most malicious attacks on watermarking algorithms based on time domain, but this algorithm showed a good performance after that kind of attack as well.

5. CONCLUSION

Based on widely employed spread-spectrum watermarking, we have developed a new algorithm for embedding and extraction of watermarks in digital audio sequences. The algorithm does not use commonly performed transformations to and from frequency domain. Experimental tests showed that the watermark was embedded in accordance with HAS properties, using temporal and frequency domain masking techniques. Detection of the watermark is done by blind watermark detection, without using the non-watermarked audio. The algorithm has high detection results, after MPEG layer-3 compression at 32 kb/s and AAC compression at 48kb/s, and against other commonly applied signal processing attacks.

AKNOWLEDGMENTS

The work is part of the Duchess project, which is supported by the Finnish National Technology Agency, TEKES.

REFERENCES

[1] P. Bassia, I. Pitas. "Robust audio watermarking in the time domain", *IEEE Transactions on Multimedia, Vol.3, No 2.,* pp. 232-241, June 2001.

[2] Mitchell D. Swanson, Bin Zhu, Ahmed H. Tewfik. "Robust audio watermarking using perceptual masking", *Signal Processing, Vol. 66*, pp. 337-355, 1998.

[3] D. Kirovski and H. Malvar. “Robust covert communication over a public audio channel using spread spectrum”, *4th International Information Hiding Workshop,* Pittsburgh, USA, April 2001.

[4] Christian Neubauer, J. Herre, K. Brandenburg. "Continuous steganographic data transmission using uncompressed audio", *Information Hiding Second International Workshop*, Portland, USA, April 1998.

[5] Jaap Haitsma, Michiel van der Veen, Ton Kalker, Fons Bruekers. "Audio watermarking for monitoring and copy protection", *ACM Multimedia Workshop Marina Del Ray, USA*, pp.119-122, 2000.

[6] W. Bender, D. Gruhl, N. Morimoto, A. Lu. "Techniques for data hiding", *IBM System Journal, Vol. 35*, pp. 313-336, 1996.

[7] ISO/IEC IS 11172, Information technology – coding of moving pictures and associated audio for digital storage up to about 1.5 Mbits/s

[8] N. Cvejic, A. Keskinarkaus, T. Seppanen, “Watermarking of audio using m-sequences and temporal masking”, *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, October 2001, accepted